the intercept project

a cluster challenging PCI-X and more…

March 6th, 2003

Supercomputing Systems AG (SCS)

Supercomputing Systems

- founded in 1993
- ~50 engineers
- up to a dozend students
- privately owned
- ~40% of turnover in HPTC

Mission Statement

Supercomputing Systems

- SCS is an independent engineering company and sells experience, not products.
- We develop customer specific (high performance) computing solutions on contract basis.
- Our solutions are mainly used in capital equipment rather than in consumer goods.

projects at SCS

Supercomputing Systems

- foto finishing system for AGFA
- control system for an electron accelerator at PSI, Switzerland (500 control processors)
- rice grain sorting; 100k / s
- Primary Flight Computer (PFC) for a vector controlled aircraft
- control system for weaving machine (10k – 100k threads)
- trigger processor for H1 @ DESY - fishing the most interesting 100 out of 10M events / s

1.1 presentation outline

Supercomputing Systems

who is SCS

1. outline
2. requirments
3. cluster overview
4. network hardware
5. communication software
6. management software
7. the present testmachine at SCS
8. results
9. outlook & discussion

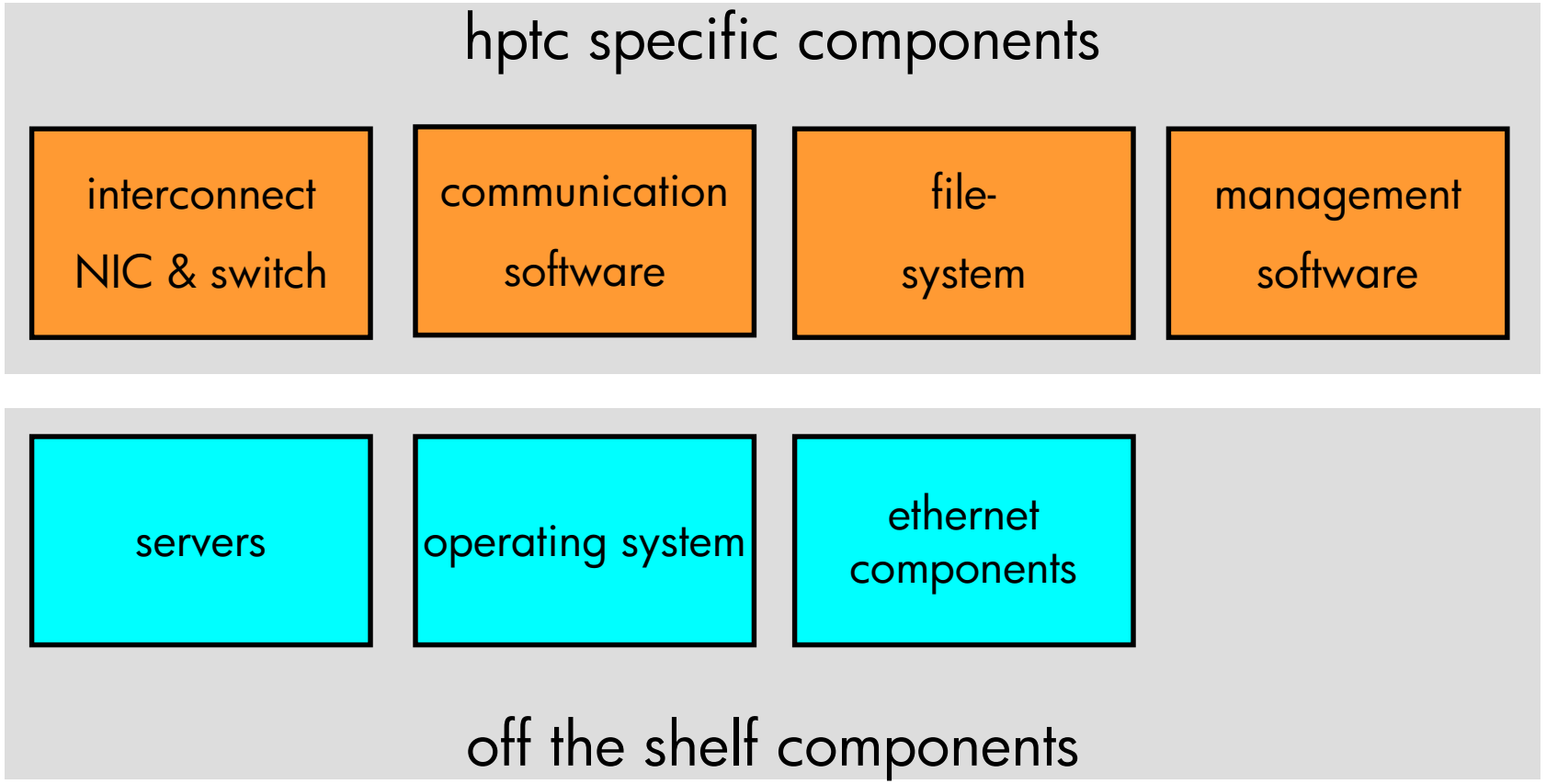
1.2 project goal

Supercomputing Systems

design and develop all components to build a supercomputer, except for the ones you can get ‚off-the-shelf'.

1.3 components to build a supercomputer

Supercomputing Systems



commodity supercomputer requirements

Supercomputing Systems

1. processor independence (IA32, IA64, alpha,...)
2. OS independence (Linux, Unix, WinXX,...)
3. no OS modifications
4. scalability up to 10k+ nodes
5. MPI
6. support for 2, 4, 8 processor SMP nodes
7. integrated management system (single system view)
8. support for any network topology (fat-tree, nD-mesh, torus, multi-stage)
9. bandwidth: 1+GB/s
10. latency: < 4 us
11. scaling I/O
12. high availability → fault tolerance
13. good cost / performance

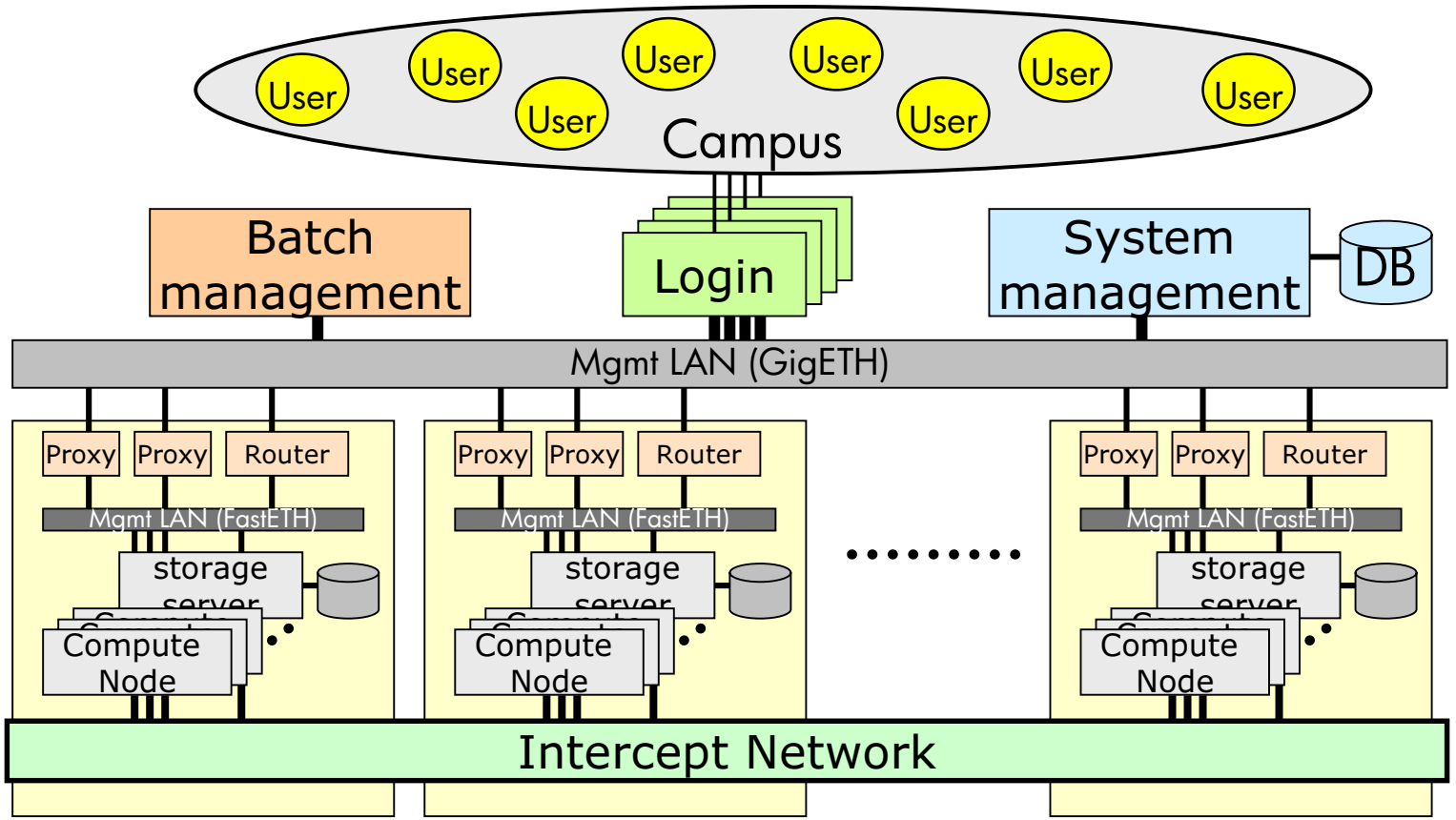
project constraints

Supercomputing Systems

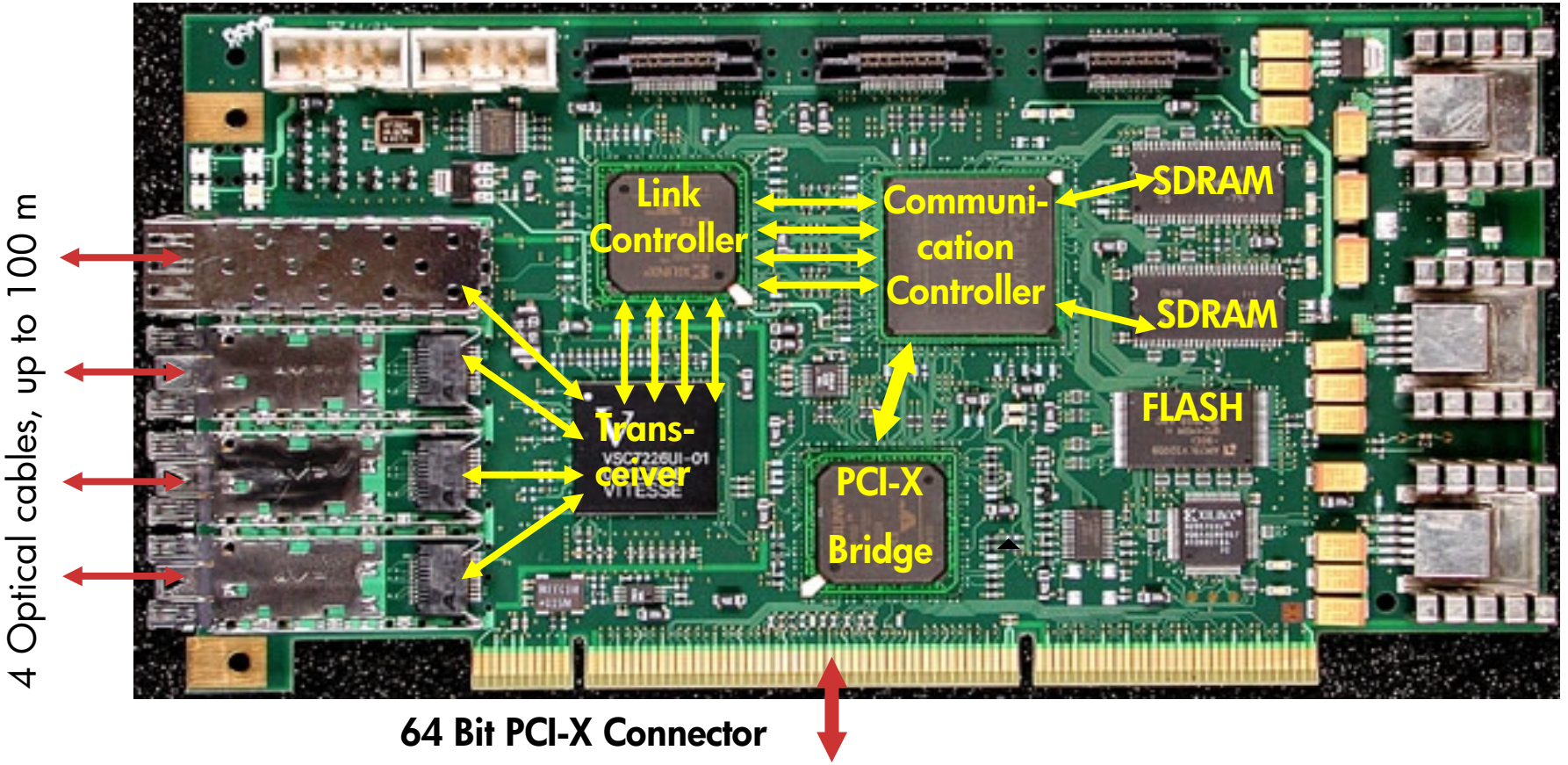
limited development effort

3.1 intercept system overview

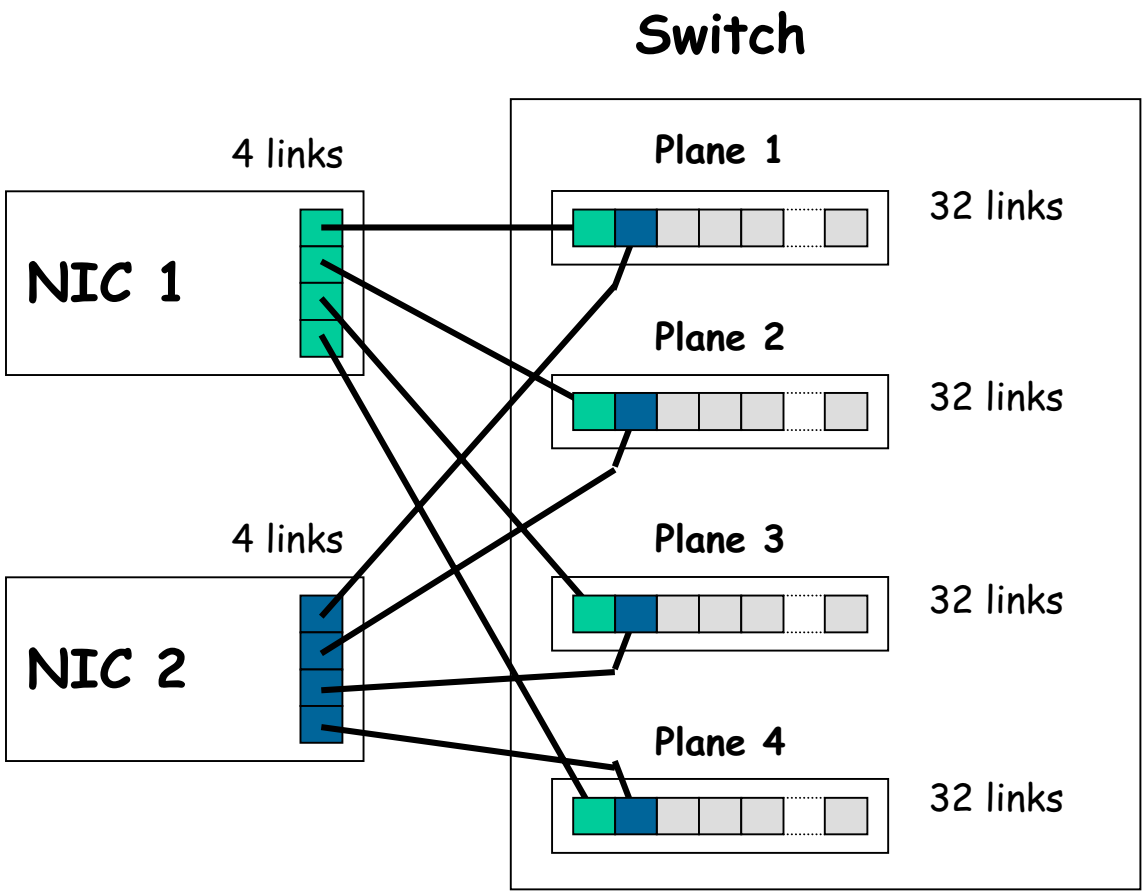
Supercomputing Systems



4.1 NIC for PCI-X and PCI (66 MHz)



4.3 multiplaning



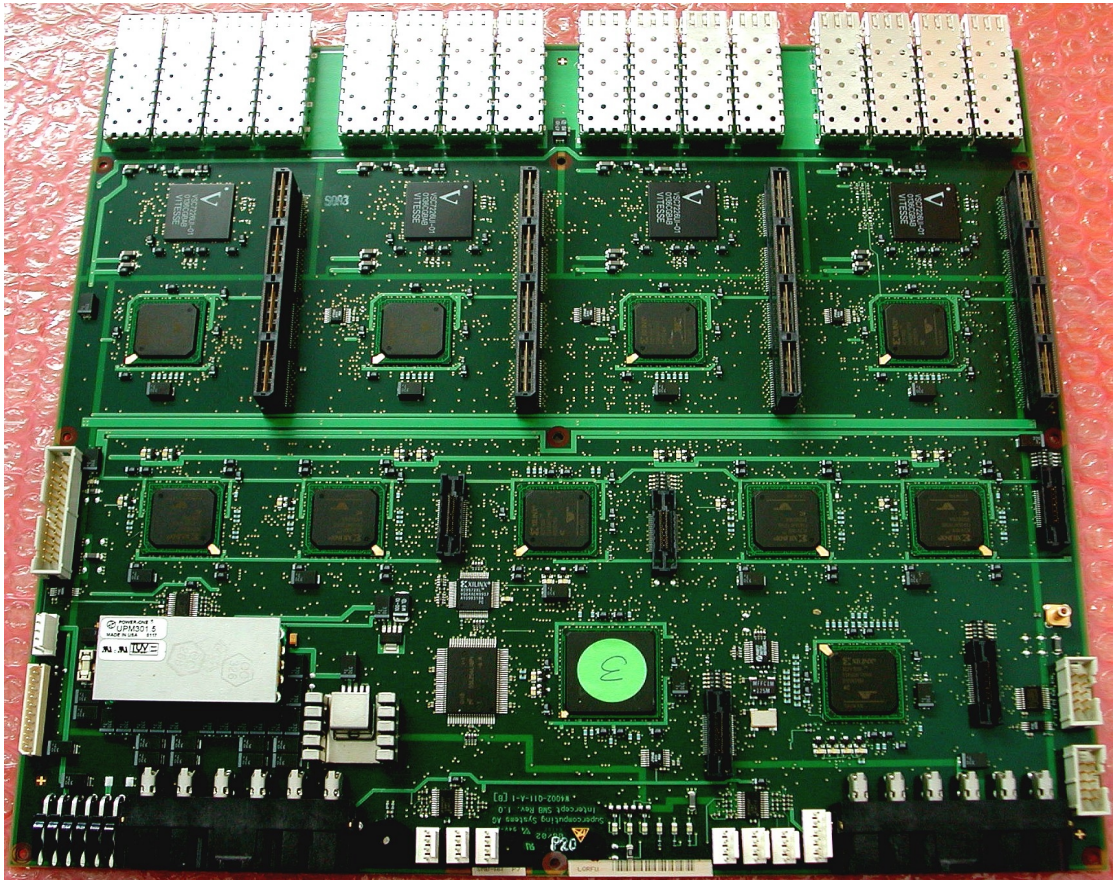
1+1=2 Gbyte/s network bandwidth

full functionality

4.5 32 port switch hardware

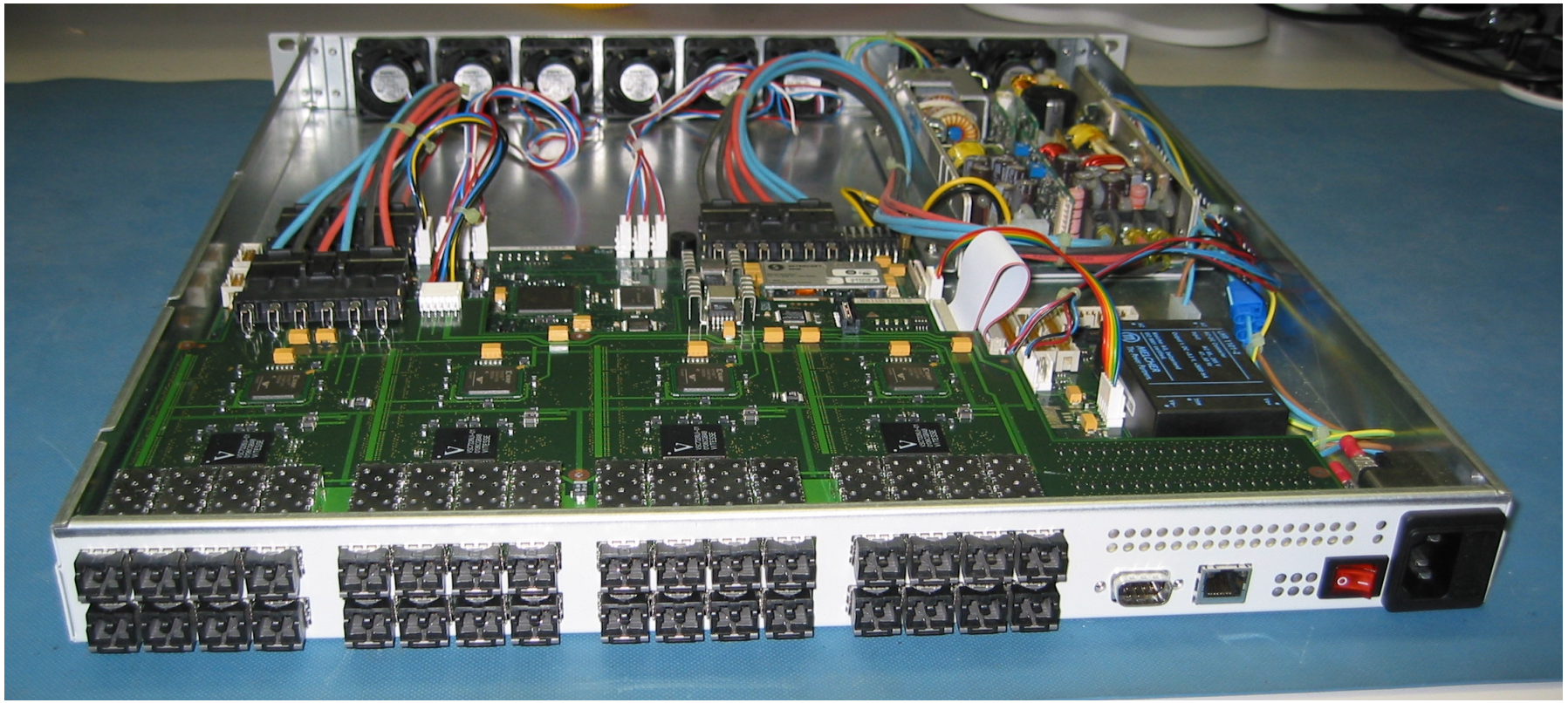
Supercomputing Systems

- Switch Main Board (SMB)
- 16 links
- full bandwidth 32x32 port crossbar
- cut-through routing
- large SRAM for routing tables



4.7 32 port switch fully assembled

Supercomputing Systems



5.1 software structure

Supercomputing Systems

component mgmt & survey through IMS

integrated batch scheduling (LSF)

MPI 1.2 (two sided communication)

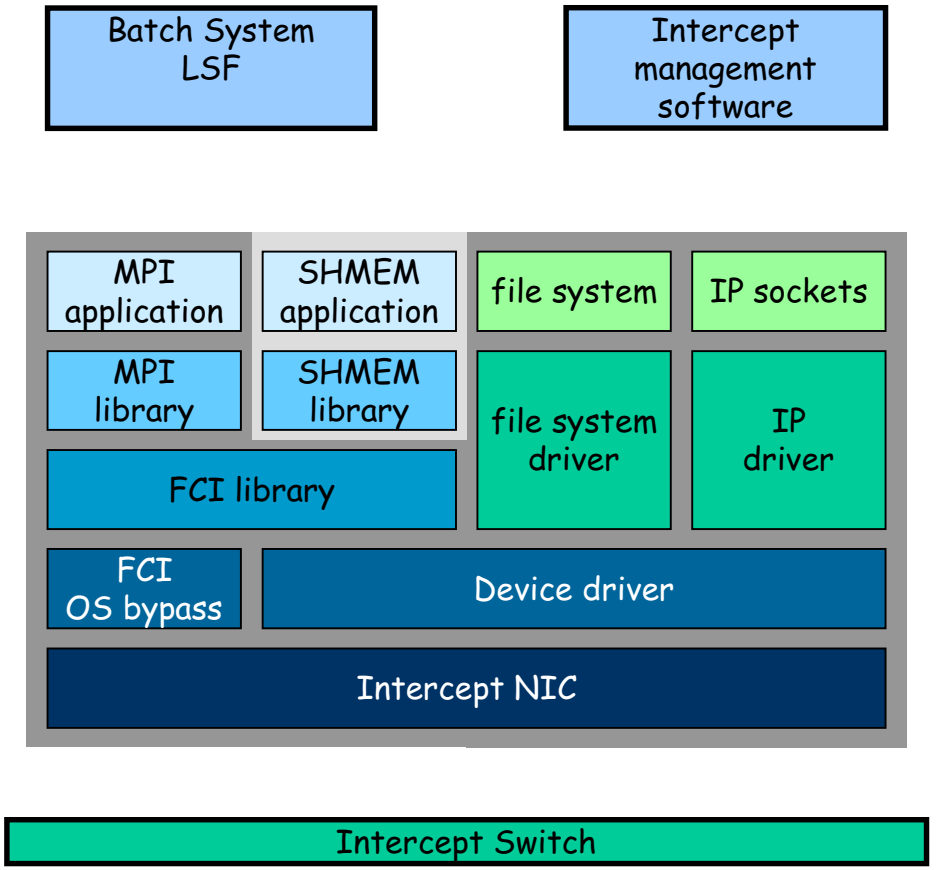
MPI one-sided communication

MPI-I/O

parallel distributed file system integration

TCP / IP

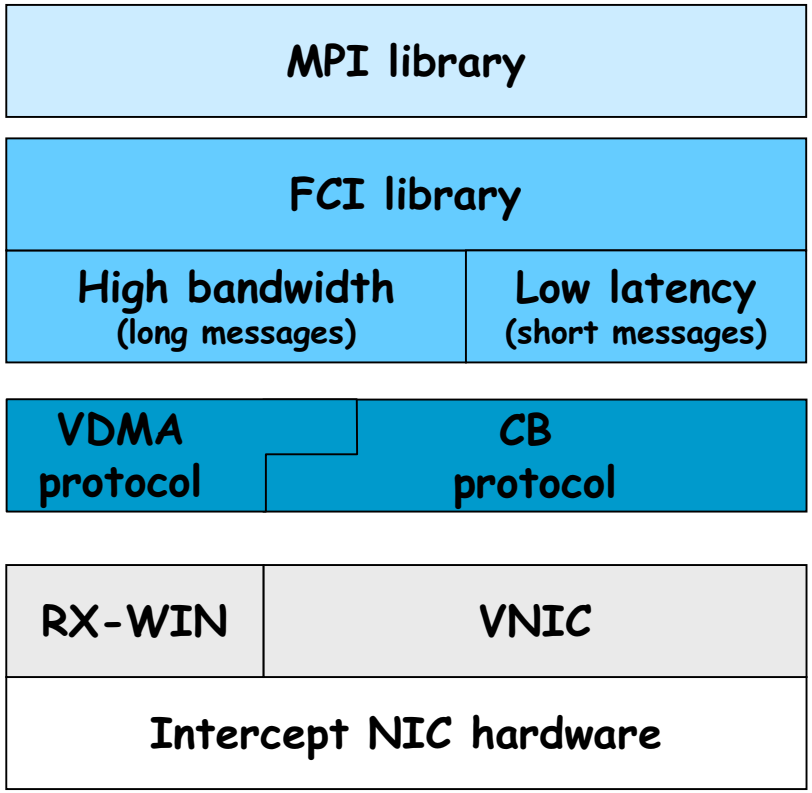
link-to-link protocol implemented in FW (send and forget)



5.2 two communication protocols

Supercomputing Systems

- two requirements
  - » low latency requires data to be sent as quickly as possible
  - » high bandwidth requires direct memory to memory transfer
- FCI makes the decision which protocol is used based on the message length
- **The only thing the user sees is the optimal performance!**

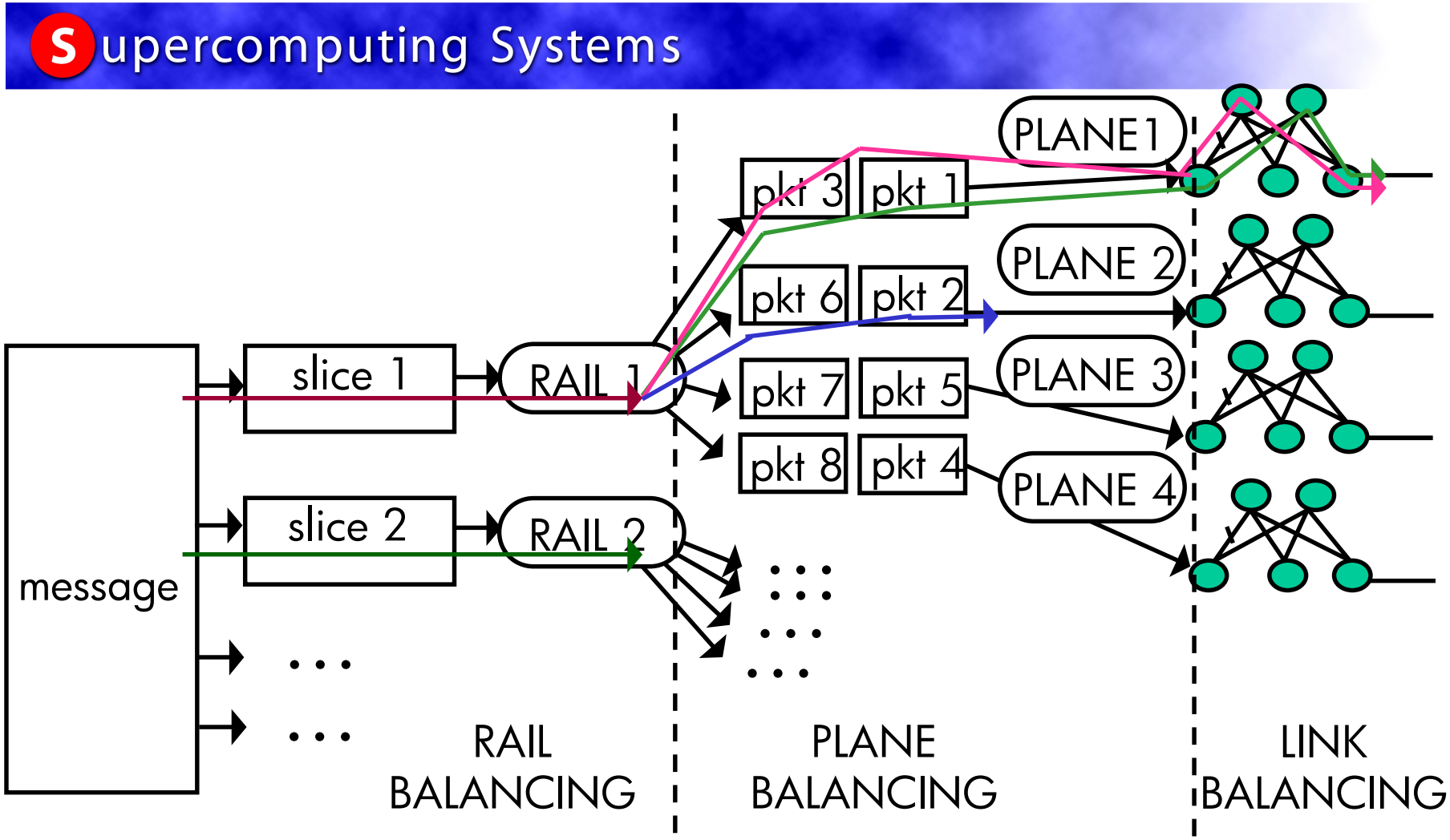


5.3 communication protocols

Supercomputing Systems

- two orthogonal requirements
  - lowest latency for short messages
  - full bandwidth at minimal processor (memory) interference for long messages
- two different protocols
  - low latency protocol for short messages and commands using circular buffers in main memory
  - VDMA protocol moves data directly from user space to user space without intermediate copy step. DMA engine on the NIC
- two sided as well as one sided communications use both protocols
- the decision on which protocol to use is taken in the communication software; users do not have to care.

5.4 dynamic routing overview



5.5 network properties

Supercomputing Systems

- reliable, flexible interconnect for high performance clusters
- full PCI-X bandwidth (~ 800 MB/s on MPI, depending on chipset)
- very low ping-pong latency (< 4 µs on MPI, 2.5 µs m2m)
- enough parallelism for 8 (16) concurrent accesses (VNICs)
- very cost effective optical solution
- very efficient MPI 1.2 and MPI One-Sided (MPI 2.0)
- optical cables up to 100 meters
- file system integration
- fat tree
- fault tolerance at all levels
- scalable design up to very large clusters; 65k nodes theoretical limit

6.1 management software tasks

Supercomputing Systems

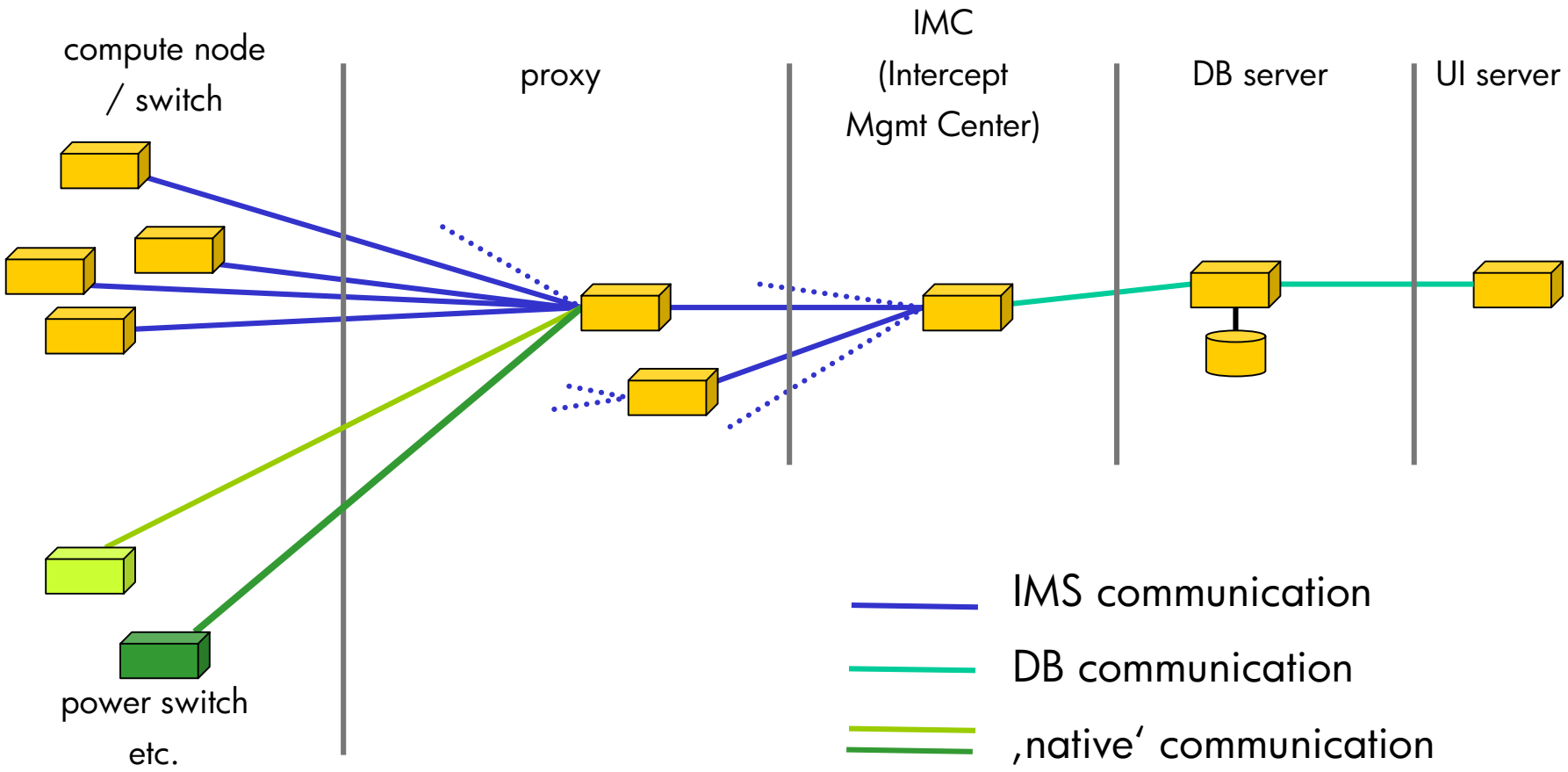
- batch management integration (LSF from Platform)
  - » provide system setup information to LSF
  - » cleans up after an application
- system installation/configuration/boot
  - » single system image due to common root file system
- network routing
  - » in case of a failure the intercept management software reroutes broken connections
- network connection analysis at bootup
  - » the correct connection is checked
- fault management
- interconnect performance monitoring

6.2 management software requirements

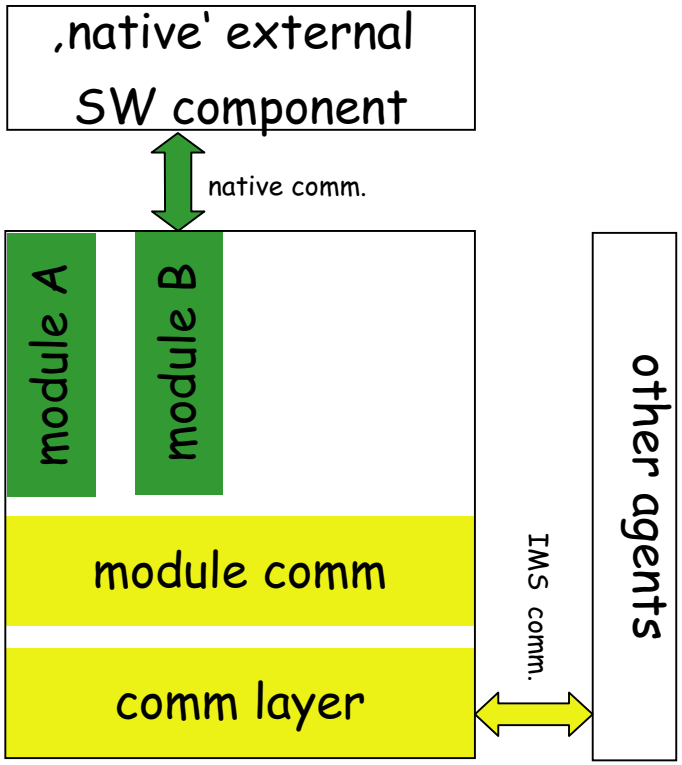
Supercomputing Systems

- scalability: proxy structure
- user friendly
  - LSF, known and widely accepted
  - graphical & command line interfaces
- portability (IA-32, IA-64, Alpha…; linux, unix, windowsXY,…)
- efficiency
  - fast (parallel) booting by using proxies
- reliability
  - data is stored in transactional database

6.3 scaling management structure



6.4 intercept management software agent



agent can be

- management center
- proxy
- node agent
- switcht agent

all agents have the same commuication layers but

- different number of modules
- different modules
- different ‚native' links

7.1 test machine at SCS (Technopark ZH)

Supercomputing Systems

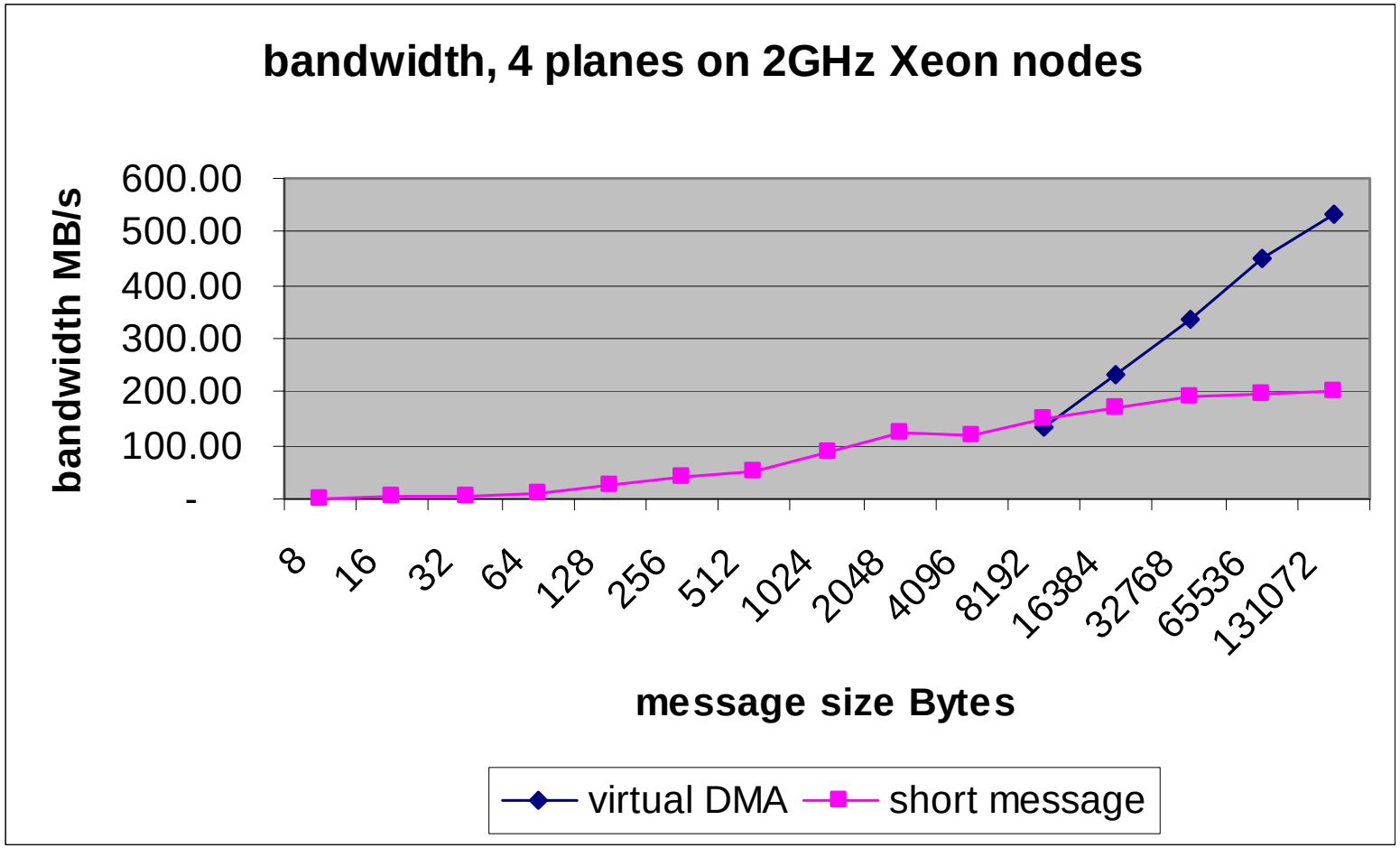
8.1 present status

Supercomputing Systems

- since november 02 the network is running on PCI-X
- MPI ping-pong latency < 4 us
- MPI ping-pong bandwidth > 700 MB/s (see later slide)
- full intel MPI testsuite run successfully
- LSF batch scheduling integrated into IMS

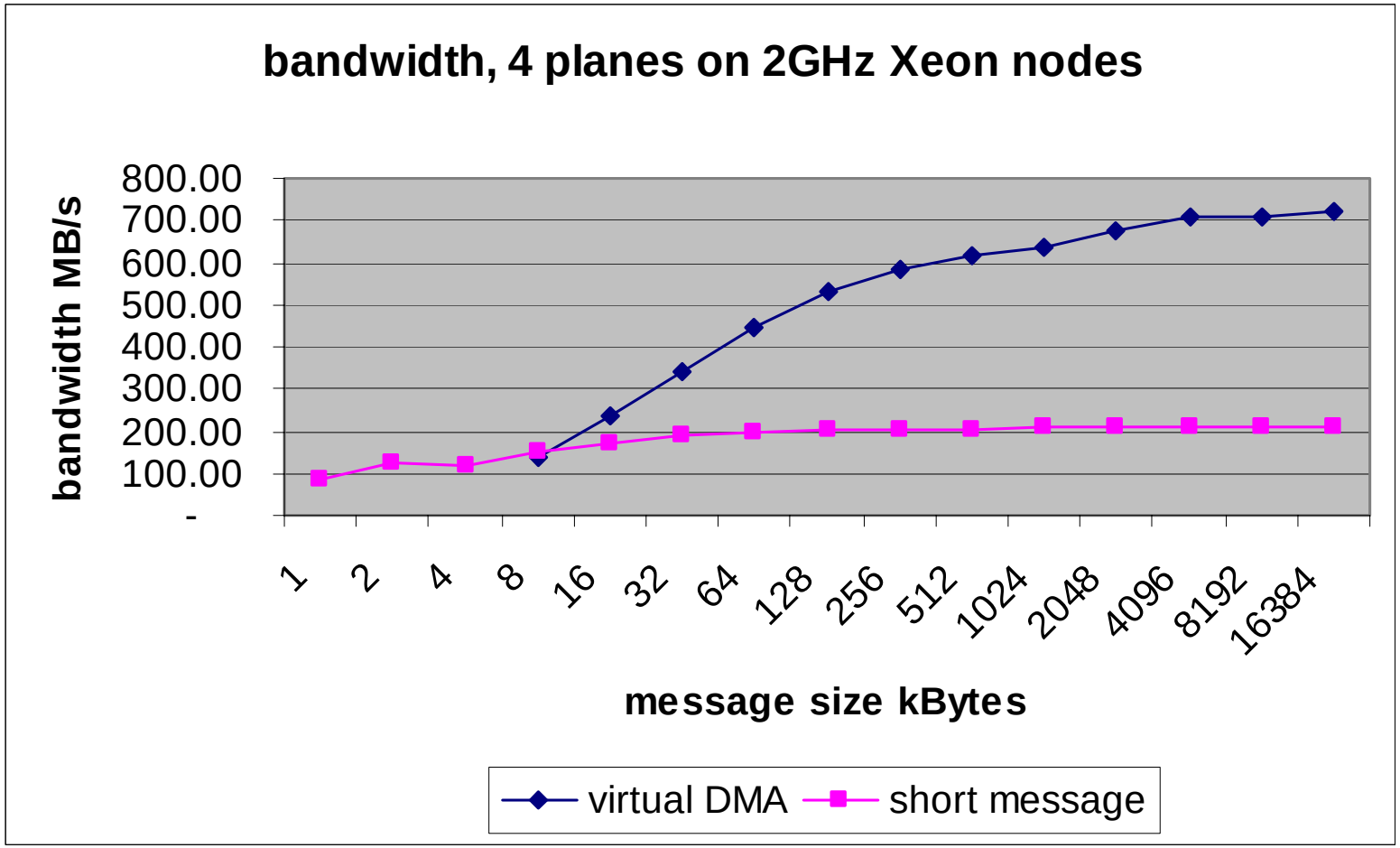
8.2 mpi ping-pong BW vs. message size

Supercomputing Systems



8.3 mpi ping-pong BW vs. message size

Supercomputing Systems



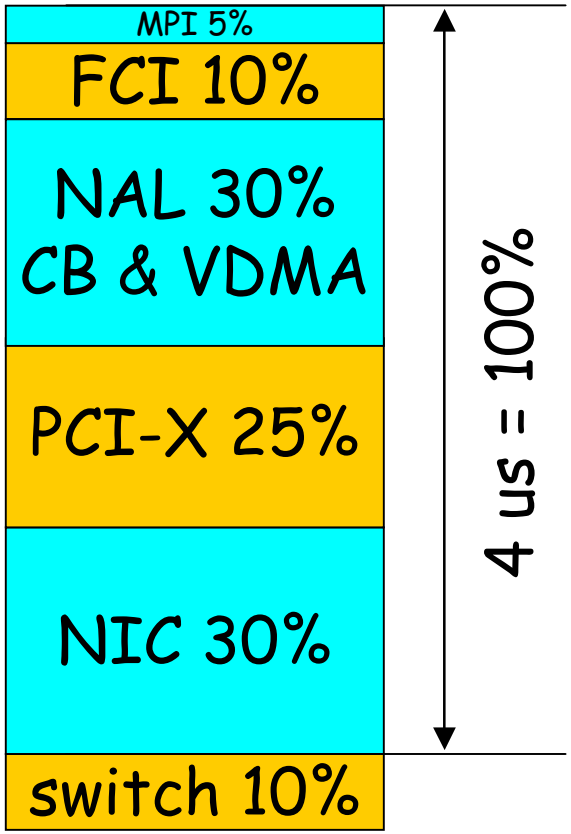
8.4 PCI-X bandwidth limitations

Supercomputing Systems

- measured on 2GHz Xeon, Serverworks GC-LE chipset
- protocol optimized with master read/write accesses from NIC to memory
  - 4kB / burst read
  - 512B / burst write
- 720 MB/s with unidirectional setup, limited by read latency
- 890 MB/s with mixed read/write (concurrent send and receive)

8.5 Latency decomposition

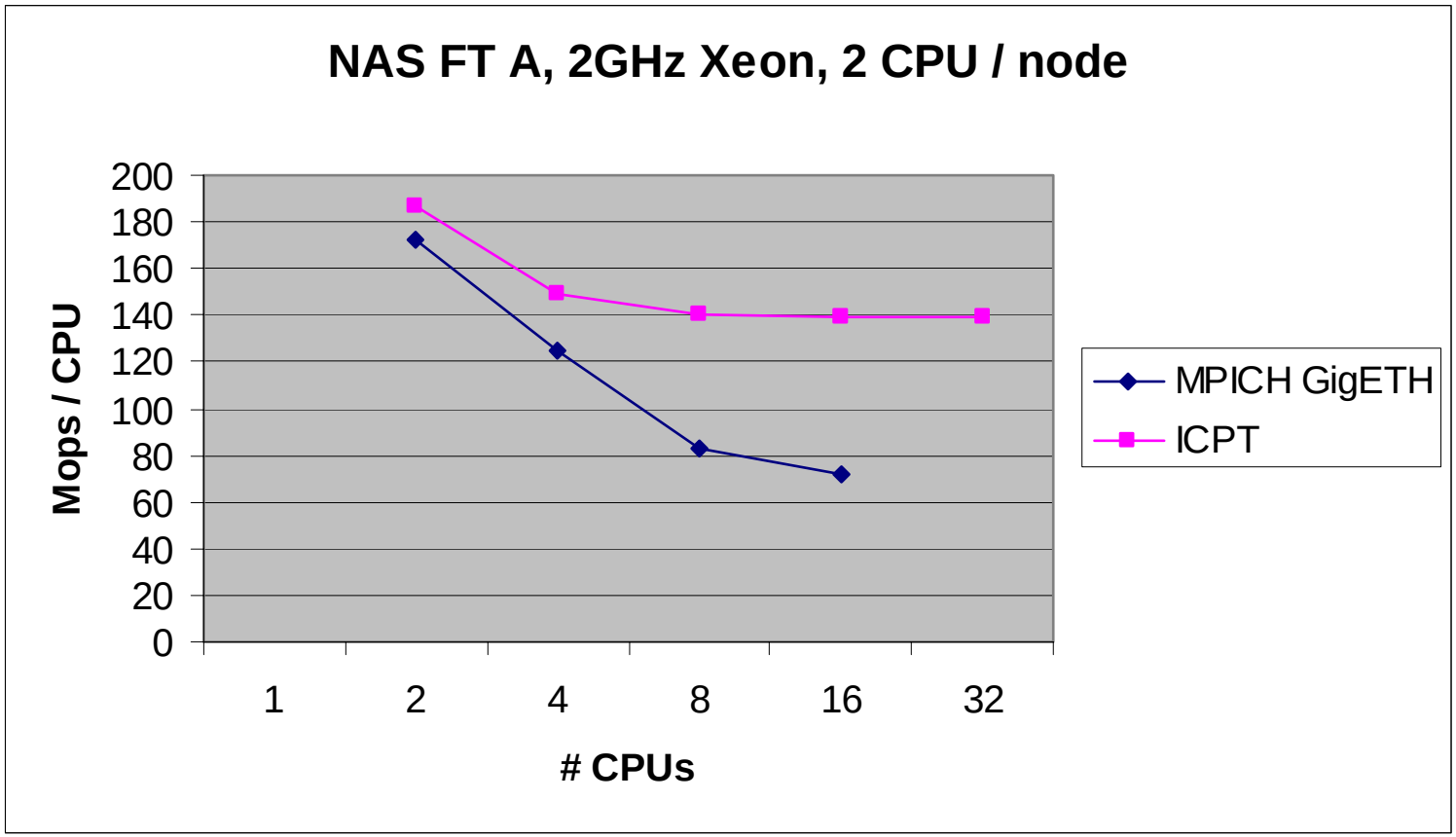
Supercomputing Systems



- NIC2NIC <4us (measured)
- normalized as 100%
- decomposition simulated
- switch adds ~400ns per hop
- FCI/MPI overhead ~15%
- PCI-X should add 25%
  (present measurements indicate that this contribution is higher)

NAS FT A per CPU performance 2 CPUs/node

Supercomputing Systems



next milestone

Supercomputing Systems

- build a large machine
- build it this summer, reach full production mode in autumn
- target performance: > 1 Tflops
- well, that's reasonably large...
- negotiations with ETH Zurich are progressing
- user applications are beeing benchmarked at SCS right now
- CPU nodes will be dual Xeon or Itanium

9.1 ‚processor on NIC' vs. ‚intelligent hardware'

Supercomputing Systems

	processor on NIC	intelligent hardware
design complexity	high	medium
software for NIC	needed	not needed
debugging complexity	high	medium
execution	serial	parallel
speed	medium	high
load on main CPU	low	low
flexibility	medium	low (ASIC) high (FPGA)
	today	tomorrow

9.2 ASIC vs. FPGA

Supercomputing Systems

	ASIC	FPGA
time to market	slow / bad	fast / very good
field upgradability	none	very good
power consumption	low / good	high / bad
achievable clock speed	>> 250 MHz	~ 250 MHz
development & initial cost	high	medium
production cost (volume)	low	high